Natural Language Processing Challenges in HIV/AIDS Clinic Notes

Sookyung Hyun, RN, MS, [1,2] Suzanne Bakken, RN, DNSc,[1,2]
Carol Friedman, PhD[2], Stephen B. Johnson, PhD[2]
[1]School of Nursing and [2]Department of Biomedical Informatics,Columbia University, New York, NY

INTRODUCTION

In recent years, significant progress has been achieved toward increased structured data entry using standardized health care terminologies. Concurrently, the value of narrative as the clinician's rich description of the encounter and source of vital information has been reaffirmed. Natural language processing (NLP) offers a strategy for integrating these approaches to provide structured reports for further computer processing. As part of a larger project aimed at using narrative data to enrich the online medical record, we analyzed a small sample of documents in a corpus of progress notes to identify potential challenges associated with using NLP for HIV/AIDS clinic notes. We provide illustrative examples of five types of challenges.

BACKGROUND

MedLEE[1] (a comprehensive extraction and encoding system in use at New York Presbyterian Hospital) has demonstrated utility in documents such as hospital discharge summaries, chest radiograph reports, and residents' sign out notes. Preliminary assessment suggested that the HIV/AIDS clinic notes include differences in content and semantics from the types of data processed in other clinical documents.

Some important data for HIV/AIDS care are typically captured in a structured and coded format, e.g., medications, diagnostic codes. However, many data crucial to providing optimal HIV/AIDS treatment are generally captured in narrative format. These include: antiretroviral regimen status and history; alternative therapies, e.g., herbal preparations; adherence behavior; signs and symptoms particularly those that are related to disease, treatment failure, or manageable side effects of medication regimens; consideration of plan of care options, e.g., alternate possibilities for antiretroviral regimens; and patient counseling and education, e.g., risk behavior, adherence strategies.

METHODS

We created the corpus by randomly selecting and transcribing one clinic note from each of the charts of 707 persons with HIV/AIDS. Perl script was used to transform soap notes into structured sections. A small sample was then processed to identify potential issues and solutions (e.g., preprocessing to replace abbreviations vs. additions to lexicon) prior to analyzing the corpus using NLP.

RESULTS

We identified preprocessing, lexical and grammar issues (Table 1). "K" was recognized for it's more common meaning of potassium rather than as 1,000 in the instance of viral load. Frequently used words in HIV/AIDS care such as "antiretrovirals" and "herb" are not currently in the MedLEE lexicon. We also identified issues common in other narratives such as zeroing, connectives, and distribution.

Table 1. Examples of Issues

Issues	Examples
Symbol not recognized or misinterpreted Abbreviation unknown	•Neurology – Cranial nerves V-VII intact. •With initial good response with viral load to 900K. •Stool O&P and culture.
Unknown lexeme - lack of recognition of a word as a lexical unit.	•Taking Herbs and Vitamin C. •Will start combination RX with above antiretrovirals. •No E-A changes.[a]
Zeroing - context required is contained in previous statement	•(Abdominal: distended but not tense.) Nontender.
Connective - component words recognized, but not joined together by connectives.	•Diarrhea ? exacerbation irritable bowel versus UTI (often presents with GI symptoms in patient) versus enteric process. •Headache Problem secondary to Foscarnet and also sinus congestion.
Distribution - modifier not distributed across a conjunction.	•Associated with shooting pain/tingling to dorsum of hand and left middle index fingers.

[a]Pulmonary auscultation finding.

DISCUSSION

Extensions to the MedLEE lexicon, which identify and categorize words and phrases, will be necessary to process HIV/AIDS notes. Issues such as disambiguation of "K" require development of new rules. Capturing connections between lists of differential diagnoses, options for plan of care, and between medications and their side effects is a greater challenge but is crucial to the understanding of the narrative. This solution will require development of complex rules to interpret the notes. Special symbols will require conversion to ASCII.

Acknowledgments

This work was supported by 1R01LM007268 and 1R01LM06274

References

1. Friedman C. A broad coverage natural language processing system. In Overhage M, editor. Proc AMIA Symp 2000: 2000:270-274.